1

Mobility and Geosocial networks

Laura Spinsanti (JRC, Italy), Michele Berlingerio (IBM Research and Development, Dublin, Ireland), Luca Pappalardo (KDDLab at ISTI-CNR, Pisa, Italy)

1.1 Introduction

The Social Web is changing the way people create and use information. Every day millions of pieces of information are shared through the medium of several on-line social networks and moreover on-line services with a social layer such as Facebook, Google+, Twitter, Foursquare, and so on. People have discovered a new way to exploit their sociality: from work to entertainment, from new participatory journalism to religion, from global to local government, from disaster management to market advertisement, from personal status update to milestone family events, the trend is to be social. Information or content are shared by users through the web by posting images or videos, blogging or micro-blogging, surveying and updating geographic information, or playing geographic-based games. Considering the increase in mobile Internet access through smartphones and the number of available (geo)social media platforms, we can expect the amount of information to continuously grow in the near future. To understand the potential of this change it is worth noticing the amount of "geosocial information" produced during recent years to be a daily occurrence. The following are just few examples. In August 2006, Flickr introduced the geo-tagging feature; by 2007, more than 20 million geo-tagged photos were uploaded to Flickr. In August 2011, Flickr announced its 6 billionth photo, with an increasing 20% year-on-year, over the last 5 years[a]. Similarly, Twitter was born in 2006. The most impressive performance indicator is the increasing rate of messages. In 2010, the average number of Tweets sent per day was 50 million[b] while

[a] Source: `http://blog.flickr.net/en/2011/08/04/6000000000/`
[b] Source: `http://blog.twitter.com/2011/03/numbers.html`

in March 2012 it has increased to 340 million[c]. In 2010, the geo-tagging feature was added to Twitter. Even considering that the amount of geo-enabled messages is only around 1 percent, this still means millions of geo-tagged messages per day. People can now be considered as sensors, producing signals on events they are directly involved in or have witnessed. Finding, visualizing and making sense of vast amounts of geo-referenced information will lead to a multi-resolution, multi-dimensional representation of the planet known as *Digital Earth*.

Such multi-modality and heterogeneity of online geo-referenced multimedia have encompassed challenges not seen in traditional geographic data analysis and mining and have attracted the attention of researchers from various communities of KDD, Multimedia, Digital libraries and Computer Vision. However, there are clearly several challenges associated with such information: the frequent changes in the data structure, the unstructured nature of contents, the limited quality control of information, varying uncertainty of geographic information and the semantic aspect on the content published, to mention a few issues. In the era of Web 2.0, the various geo-referenced media are mostly socially generated, collaboratively authored and community-contributed. The temporal and geographical references, together with textual metadata, reflect where and when the media was collected or authored, or the locations and time described by the media content. The enriched online multimedia resources open up a new world of opportunities to discover knowledge and information related to location and our human society.

Social networks that also use and create geosocial information have grown in importance and popularity, adopting names such as location-based mobile social networks, or geographic social networks, or simply social networks with geographic features. In general, there exists several types of media with temporal and geographical references on the Internet: (1) geo-tagged photos on photo-sharing websites like Flickr, (2) geo-referenced videos on websites like Youtube, (3) geo-referenced web documents, like articles in Wikipedia and blogs in MySpace, (4) geo-referenced microblogging websites like Twitter and (5) "check-in" services (users can post their location at a venue and connect with friends) such as Foursquare. Most of these services publish unsupervised (geospatial) content. Their importance has grown in such a way as to also have several definitions such as: *crowd sourcing* which consider users as sensors for gathering data; *distributed intelligence* where users

[c] Source: `http://blog.twitter.com/2012/03/twitter-turns-six.html`

are basic interpreters or preprocessors in transmitting information; *participatory science* when citizens participate in problem definition, data collection, and data interpretation; *volunteered geographic information* (VGI) when the contributive aspects is crucial; *contributed geographic information* (CGI); or just *User Generated Geographic Content* (UGGC). Some ambiguity in the use of different terms exists, such as an increasingly, however reference to crowd-sourced data in geography as volunteered geographic information (VGI) without distinguish different levels of participation (or voluntariness) when providing information. The term CGI maybe better suited to act as an umbrella term and, therefore, will be used in the rest of the chapter.

The voluminous geo-referenced contents on the Internet are a result of collective geo-tagging by the web community. Geo-Tagging refers to the process of adding geographical identification metadata to media resources, such as photographs, video, articles, web sites and so on. The metadata usually consists of latitude and longitude coordinates and, sometimes, altitude, camera heading direction, IP address and place name. In general, the means of geo-tagging can be classified into two types: integrated hardware (automatic), and purely software solutions (manual). GPS and other geolocation acquisition hardware provide an automatic solution for geo-tagging contents. However, till now, only a small portion of geo-referenced information is geo-tagged via these means and any geographic information mostly depends on the nature of the content. For example, most geo-referenced photos on the Internet are tagged by web users manually via a geo-tagging software platform. To facilitate easy geo-tagging, commercial media sharing services have adopted map based tagging tools. In general, these geo-tagging tools allow a user to drag and drop photos to a location on the map. The intuitive map and user-friendly interface render the geo-tagging a simple and straightforward process. However, the major limitation of such geo-tagging processes is that, currently, no industry standards exist on tagging and storing the geo-tags of media. Most commercial media repositories store geo-tags in tag-based systems, similar to how text tags are stored. The most important consequence is that several facets of uncertainty are related to the location that can be retrieved as we describe later in the chapter.

The rest of the chapter gives an overview of existing and foreseeing applications that use this CGI data with particularly focus to mobility. It then describes the problem to reconstruct trajectories and the research issues related to geographic and semantic uncertainty of this

data. Several open issues still remain due to the novelty of this research area.

1.2 Geosocial data and Mobility

The use of geosocial data covers a wide range of possible applications, essentially all the contexts in which the location (and time) play an important role such as health, entertainment, work, personal life, tourism, etc. While we want to focus on the mobility aspects of geosocial data, this topic is really a forefront research of latest years. The studies conducted so far have started based on several works produced on mobile phone data. Despite the similarities of this data, as explained in section 1.4.2, the conceptual framework and the characteristics of geosocial data leads to a real new branch of research. The researches about this new domain are far off from being exhaustive. As described in section 1.3 trajectories resulting from geosocial data are built from a collection of sparse data points. This ends up in different groups of applications, as described below.

We can distinguish a first group of applications that use only the location from geosocial data, generally to filter the contents (message, photo, video, news, tweets and so on) from a zone they want to analyze or they want to receive alerts (newsfeed mechanism). For example, the impact of (geo)social media during crisis events has shown the high value for relief workers or coordinators, and the affected population. Examples from the natural disaster field include wild fires in the United States and France, hurricanes in the United States, the 2010 earthquake in Haiti, and floods in the United Kingdom, while an example from social-political field is the Arabic revolutions started in late 2010. In all these cases, messages were filtered using the related location such as coordinates, user location settings or place names in text or tags.

Another group of applications use the set of places to discover patterns. An example is the tourism knowledge scenario. In Web 2.0 communities, people share their travelling experience in blogs and forums. These articles, named travelogues, contain various tourism related information, including text depiction of landmark, photos of attractions and so on. Travelogue provides abundant data source to extract tourism related knowledge. Travelogues can be exploited to generate location overviews in the form of both visual and textual descriptions. The method consists first in mining a set of location-representative keywords from trav-

elogues, and then in retrieving web images using the learned keywords. The model learns the word-topic (local and global tourism topic, like an attraction sight) distribution of travelogue documents and identifies representative keywords within a given location. Complementing travelogues, geo-referenced photos also tell a great deal about tourism knowledge. The photos, together with their time- and geo-references, implicitly document the photographers spatio/temporal movement paths. The tourist visited points can be grouped, mined to distinguish patterns and used to rank places of interest and generate recommendations. In most of these cases, applications use location extracted from people trajectory in the real world, but they are not really using the trajectory itself.

A third group of applications also considers the user interaction and relationship. In fact, geosocial networks provide not only the location, but also the explicit social links, and in some cases explicit declaration of kinships and partnerships, giving the possibility to overcome the shortcomings of techniques to infer tie strength. They also give high resolution location data, as one can distinguish between a check-in to different floor of the same building. To give an example, Yahoo! research labs published a study on the attempt to extract aggregate knowledge on certain locations from large scale geo-referenced photos at Flickr. The knowledge here refers to the word or concept that can best describe and represent a geographical region. The challenge is to extract structured knowledge from the unstructured set of tags. The premise of the proposed solution is based on the human attention and behaviour embedded in the photos and tags. Namely, if tags concentrate in a geographical area but do not occur often outside that area, then these tags are more representative to the area than those spread over large spatial region. This example shows also that there is a need to model human behaviour and this aspect constitutes an interesting research topic by itself. Of course, models and hypotheses are geographically dependent as western people often act differently from eastern people in social context. However, online social networks check-ins are usually more sporadic than phone calls, providing less temporal resolution than mobile data.

Some references for further lecture are provided in the annotated bibliography section.

1.2.1 Foreseeing applications scenario

In this section we describe some possible scenario where the use of geosocial data can be useful, but not yet investigated by researchers. The anal-

ysis of virtual movements in geosocial networks could be useful in several application scenarios. For instance, in the emerging field of human dynamics, a central point is the understanding of the interplay between human mobility and social networks. How do the mobility patterns and parameters depend on social network characteristics? The study of such interaction needs massive society-wide datasets that simultaneously capture the dynamical information on individual movements and social relationships. Traditionally, this problem is addressed by using mobile phone networks, because they provide at the same time temporal information and social contacts. However, there are at least two problems with this kind of mobile phone data. Firstly, friendships are not explicit but are inferred by creating a who-called-whom graph, with the possibility of inaccurate information about tie strengths. For example a person does not call so often people who live with him. The low number of calls between them is interpreted as a weak tie, leading to a bad representation of reality. Secondly, we know users' positions only when they perform a call, and merely we know the position of the tower managing the area the user is within, and not the actual geographical location of the user. The use of geosocial data to this domain can lead to more accurate results.

The spreading of biological and mobile phone viruses is another context in which geosocial data could be useful, because epidemics are also determined by the structure of social and contact networks within the population, and the human mobility patterns of people. The mathematical modelling of infectious diseases must take into account travel patterns within a city or the entire world, and accurately shape the underlying contact network depending on the nature and the infectiousness of the pathogen. For example, with highly contagious diseases (e.g. transmission based on coughs and sneezes) the contact network will include any pair of people who sat together in the same place. For a disease requiring close contact (e.g. sexually transmitted disease), the contact network will be much sparser. Similar distinctions arise in computer viruses context, where a malware infecting computers across the Internet will have a much broader contact network than one that spreads by short-range wireless communication between nearby mobile devices. Depending on the case, a contact network based on co-location in a place or the explicit social network could be inferred by using geosocial data from geosocial networks, geo-tagged photo websites or geo-referenced microblogging websites. Some researches in this direction have been conducted, but far away to be exhaustive.

Mobility patterns of a population can be extracted by using check-in

trajectories of users, in order to define the epidemic model or to perform a simulation scenario. A very fascinating application is the developing of mobility models and routing algorithms for the so-called opportunistic networks. They are a new paradigm of computation in which there is no a fixed infrastructure, and mobility is exploited as an opportunity to deliver data among disconnected parts of a network. When a node has data to transfer toward another node, and no network path exists between the sender and the receiver, any possible encountered mobile device represents an opportunity to forward and carry them until encountering another node deemed more suitable to bring the message to the eventual destination. Both in the design of routing algorithms and in the evaluation of them, a promising approach is that of incorporating spatial dimension into a model based on time-varying social graphs. Geosocial data are clearly the most appropriate and useful tool in this context because provide at the same time all the three dimension of human movements: spatial, temporal and social dimensions. In addition to this, explicit social relationships from online social networks can be incorporated to better design protocols that are able to learn the social network of users, for example in order to exploit the role of hubs (users with the highest number of contacts) in the dissemination process, or to predict new friendships and contact opportunities.

These examples are, of course, not exhaustive. They just give a hint of the possible uses of CGI data in several different context scenarios. The technology and the application are moving and changing so fast that some very unexpected and innovative application can be developed even in the next few months.

1.3 Trajectory from geosocial web

Users in the social web leave footprints of their movements: they visit real and virtual places and their movements can be recorded and analyzed. Following the previous paragraph scenarios, we want then answer the following question: Which kind of trajectories can we reconstruct from the geosocial web data?

Except from an experiment carried out by Microsoft research with GeoLife project in which 165 users tracked their GPS trajectories on a social platform, the data we can commonly retrieve and access from geosocial networks is punctual and discontinuous. The main reason is not related on the location systems, but on the users communication behaviours on

social networks. Generally a user posts a content when it is important for her to share with others users, or friends. This means that he is not interested in communicate in a continuous way. Moreover some media are more used in specific circumstances (i.e. photo sharing/repository during holidays), while others in daily routine (i.e. check-ins style or status update). Follow a single user on a single social network generate a finite list of spatio-temporal positions, that can be used for implement discrete trajectory (see Chapter 1). This use of discrete position is in a very early stage and has still several limits. One example is Google increasing popular service called Google Latitude, that allows users to share their location with friends and add it to their status message in other Google applications. The history option (in a beta release at the moment of writing), stores the users past locations and visualizes them on Google Maps and Earth. The user (and only she) can visualize the trajectory and a dashboard showing information, such as trips, frequently visited locations, distance travelled and time spent in different places. This application uses raw data to reconstruct the user trajectories and enrich them with semantic information, as described in Chapter 1 for semantic trajectories and behaviours. In Figure 1.1 is possible to see one of the authors Google Latitude trajectory from one month data. As is possible to see, the trajectory reconstruction in the social network has some challenging issues as, for example, the data acquisition, given that the data can be discontinuous in time. In the figure is possible to see long straight lines connecting far points on the map. There is no attempt to connect to road map layer or transportation means.

Following a single user in his daily social networks activity on different social platforms could help in creating different trajectories or in filling some gaps with respect of using only one social network source. A very nice example of a segmented trajectory (see Chapter1) is shown in Figure 1.2, extract from an advertising of a train Wi-Fi connection. The option to share information between different social networks publishing contents from one platform to another platform is a very recent trend and it is not yet studied by the scientific community.

1.4 Geographic information in Geosocial Web

In the following sections we focus on the geographic aspects related to information it is possible to retrieve from the Social web. We then answer the following questions:

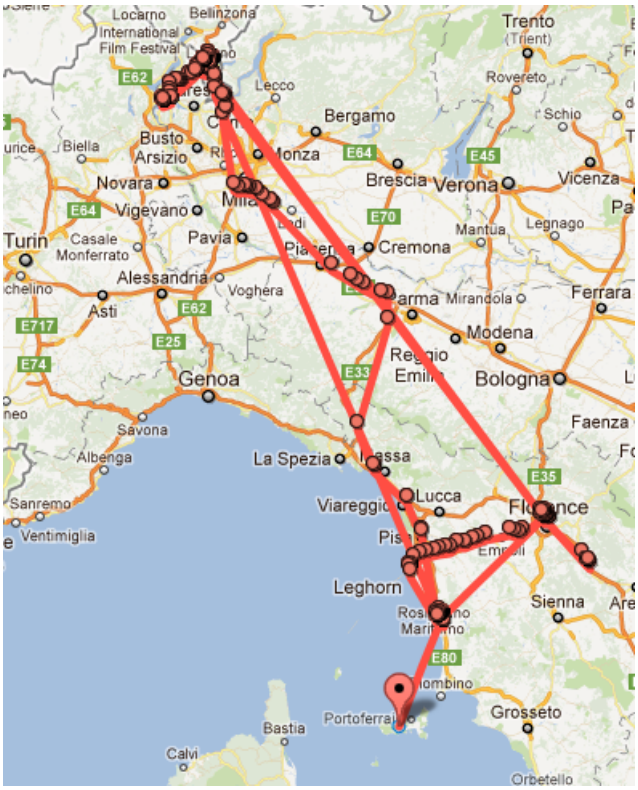

Figure 1.1 One month trajectory from one of the authors Google Latitude logs.

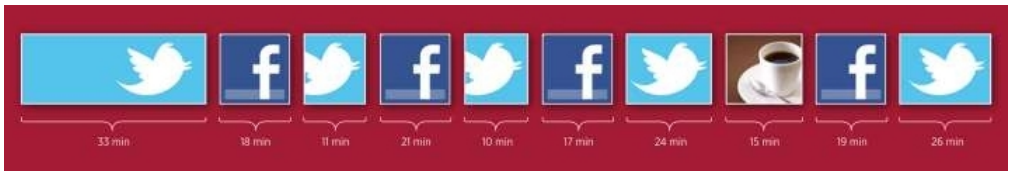

Figure 1.2 A social network use segmented trajectory.

- How does location information relate to generated information on the web? (section 1.4.1)
- How might location be captured and represented? (section 1.4.3)
- How can trajectories footprints in the web be retrieved? (section**??**ec:geo.footprint)

- What are the possible sources for uncertainty (with respect to the location information)? (section 1.4.4)

1.4.1 Location: from real world to geosocial world

We refer to content as any piece of information (such as text, image, audio and video in any possible format) that is possible to publish on the web as a resource. Content is generated by a person (that represents him/herself or a broader entity like an enterprise or an agency) using a device. The content describes a real/abstract object/event. Based on Oxford dictionary definition of real and event, we refer to real object/event as "actually existing as a thing or occurring in fact; not imagined or supposed". A real object is perdurant entity in the word such as a mountain or a building and a real event is "a thing that happens or takes place, especially one of importance" in a specific place in a limited amount of time, such as a forest fire or a football match. We refer to abstract object or event for every other information, including mood and feeling description, such as messages like "I really feel good, today". Even if abstract messages can have associated geographic coordinates, we limit the discussion to the real objects or events and we call them *Features of Interest*. In Figure 1.3 we can see three levels: the real world, the content and the social web levels and the relations among objects in the different levels. The entities in the bottom part (the real world) are the person, the device and the feature of interest. Each of them has a spatial location and extension.

Any information produced is called content. A piece of information associated to a content to describe some properties of the information is generally referred as metadata. A geographic content, or CGI, has associated *Geospatial Information* that represents a spatial reference and geometry in any format. In other words, the metadata also contain the geographic reference. The metadata can be automatically generated by the device (such as the date for a digital photo) or manually added by a person (such as the title or the tags). GPS device can record the coordinate of the device and associate the geographic information to the content metadata. Content can also include *implicit geographic information* such as place name in a textual message or the object represented in a photo. The implicit information can be made explicit using different applications and strategies and added to the content as metadata. The content with its metadata is published on the web and become *published content*: a shared resource for a certain community. At the web level it

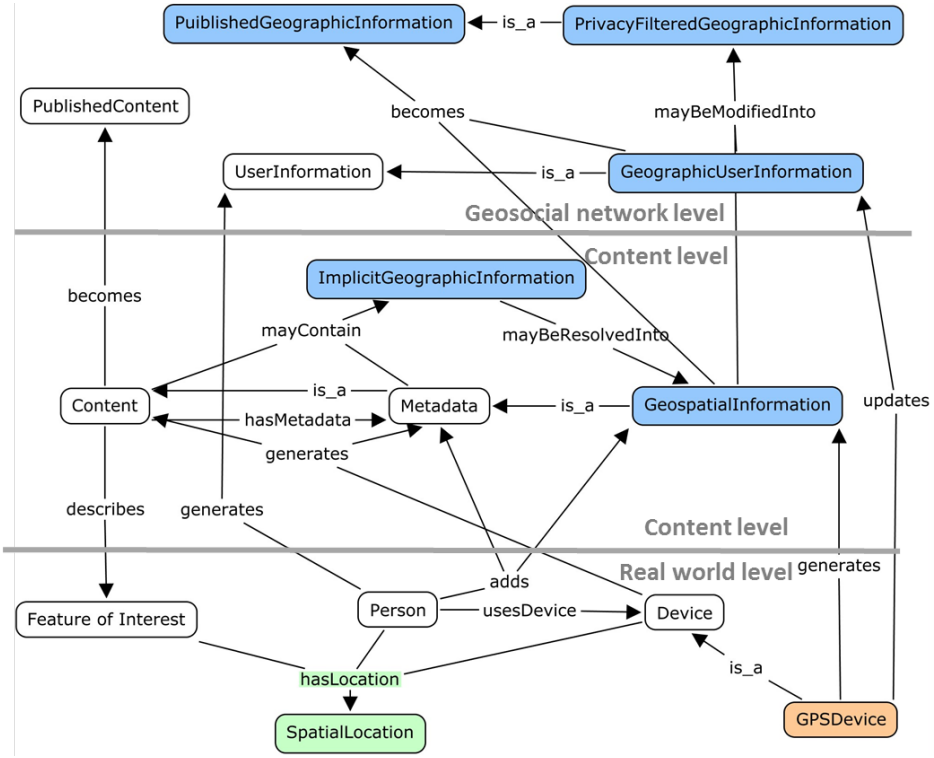

Figure 1.3 Conceptual model of CGI. The blue concepts represent where the spatial information could be retrieved, while the green one SpatialLocation represents the physical extension of the entities in the real world (person, device and feature of interest).

is also worth noticing that a person has a virtual identity. His/her personal information on the social web can include geographic information related to his/her usual living place, visited places and/or actual location. We call them *geographic user information*. This data, especially the actual position, can be manually set or automatically update from a GPS device. The geographic information contained in the different levels may coincide or not. To summarize and answer the first question posed in the introduction, spatial location associated to a feature of interest or to a person in the real world may have a representation in the geosocial network, very often associated to a content as metadata.

1.4.2 Comparison of GPS, GSM and online geosocial network data

In the last few years, many researchers studied geo-tagged data such as GPS traces, GSM data, and data coming from geosocial networks. This data coming from GSM and GPS sources look different, and may seem not even comparable with the geo-tagged data present in on-line social networks. There are, however, some aspects that can be compared, and the purpose of this section is to shed some light on differences and similarities, in both the data sources and the final tasks of analysis that can be performed on that. Let us first review the different kinds of data that we refer to. Our first source are GPS (Global Positioning System) data. There is a large variety of devices dealing with this kind of data: mobile GPS navigation systems, GPS loggers, GPS anti-theft system, GPS units for photo cameras, and so on. Clearly, the final use of them may differ, but they have the kind of information they deal with in common: they all take the global coordinates (latitude, longitude, and, sometimes, time) of the device and store it for a specific purpose. Most of them (loggers and navigators, for example) take the information periodically (every second or so, depending on the final application) and store it for the final purpose. As we have seen also in Chapters 1, 2 and 3, a typical line of such data may contain at least the following information:

```
ID, timestamp, latitude, longitude, quality of signal
```

where "ID" is the device identifier, "timestamp" is the current time, usually expressed in seconds since the 1970 (higher resolutions may be needed depending on the application), "latitude" and "longitude" are the GPS spatial coordinates, and "quality of signal" may give information on the accuracy of the measurement. Depending on the application, a user may personally produce small to large amount of data of this kind, with up to a few records per second, recorded continuously. Also, different sources of this kind of data are available, most of which are not publicly available: anti-theft systems, GPS loggers, and navigators, for example, are meant to be used for personal use, and there is no clear reason why the data they produce should be available unless the owners what to share it. Another different source of data is the GSM CDR (Call Details Record): when placing mobile phone calls, the users generate a large amount of data about their calls: number called, time, duration, and so

on. As seen in Chapter 2, a single record of this data has usually the following format and information contained in it:

```
callerID, receiverID, time, antennaID, start, stop, callID
```

where "antennaID" is the identifier of the GPS base station the phone is attached to in that moment, and "callID" is used to track the call through different antennas in the case of a user moving in space. As we see, the spatial information of this data is much less precise: while the GPS can be accurate to the centimeter, in GSM data we can only use the antenna ID as geographic information, and this is very rough. In fact, a single antenna can cover a round region of very diverse radii, depending on the power of the antenna, on the placement of it (within city, countryside,...) and other factors, giving then an estimate of the position of the caller with a precision that usually is on the order of hundreds of meters. Moreover, this is clearly sensible information, and this kind of data is usually rare, and harder to find w.r.t GPS data. Lastly, there is the data coming from the on-line geosocial networks and services. This data is of a different form, w.r.t the previous two types as, in addition to the potential geographic information contained in it, it also typically includes the content: a block of information in the form of a text message, a picture, a video, and so on. The content is said to be geo-tagged by the first block of information (i.e., the one containing the geographic information). The geographic information contained in this kind of data is usually a derivative (when not passed as-is) of the GPS data coming from mobile devices (phones, PDA, cameras) from which the user generated the content, thus we can still consider this as a GPS data source, even if, given the very particular features of the application (no need to continuously track the user, no need for a specific precision, and so on), this data is differentiable from the one coming from the GPS navigators and loggers. We will discuss later about geographic information added manually to the content. Given the large amount of geosocial networks and services available nowadays, it would be impossible to list all the possible information now available online. We can, however, present here three different types of services, representative of a large set of available online social networks: Twitter, Flickr and Foursquare. Twitter is a social network where users can post short messages in their timeline (typically publicly available), that will appear automatically within the timeline of all their followers. A typical message is a text message no longer than 140 characters, that may contain text and URLs for attached media such as pictures or videos. The messages

can also be geo-tagged if the user has enabled this feature. A typical data then contains the following information:

```
userID, messageID, text, geo-location, timestamp
```

Flickr is a photo sharing service with a social network layer where users can post pictures and video in their profile. Tags, comments, geo-location, EXIF data (technical data about the picture) are usually associated to the pictures (or videos). A typical data regarding a picture contains the following:

```
pictureID, userID, geo-location, timestamp, tags, comments
```

Foursquare is a location-based social network where users can post their current location and share this with all their friends. The service includes game features to incentive the users to share their location. A typical data contains the following:

```
userID, geo-location, locationID, timestamp
```

Source	Public	Volume per user	Precision	Sensible	Social layer
GPS	No	High	1cm	Yes	No
GSM	No	Low to high	100m	Yes	Yes
Geosocial nets	Yes	Low to moderate	1cm to 1m	No	Yes

Table 1.1 *Summary of typical properties of mobility data from GSM, GPS, and geosocial networks sources (real-world scenarios may differ depending on the application)*

Table 1.1 summarizes a few properties of the data described above. Note that they are typical properties, single examples on real-world scenarios may differ depending on the application. As we see the three sources of data differ in public availability, volume of data usually generated per user, accuracy of the data, whether the data should be considered as sensible for privacy reasons (in on-line social networks usually the data is sparse, and provided intentionally by the users, bringing this type of data to a reasonably non-sensible status), and the social dimension (exist a social connection between two users). Clearly, given the above characteristics, the tasks of analysis to be performed on the different data are very different, and each task should be conducted on the most appropriate data. When assessing, for example, the validity of

a urban transportation system, using on-line social network data may result inappropriate, as the data does not contain enough and precise information.

1.4.3 CGI retrieval from geosocial networks

All major geosocial networking systems offer access to their huge corpus of data via several Web APIs (Application Programming Interface). Many developers have created and made freely available libraries that do a lot of the heavy lifting needed to interact with the APIs, allowing researchers and data analysts to reconstruct and explore portions of social graphs and users movements. An API provides methods to access almost every feature of the system, and is typically defined as a set of HTTP request messages along with a definition of the structure of response messages (usually in an XML or JSON format). Each API is in constant evolution and represents a facet of the system, allowing developers to integrate specific functions or to build upon and extend their applications in new and creative ways. However, as regards the downloading of data, it presents some limitations due to compliance with privacy policies or the management of server load. Such restrictions define the level of detail and accuracy with which is possible to get data. We present a quick discussion of the API challenges of three very popular geosocial networks.

Twitter, for example, currently provides three APIs. Two of them offer methods to access status data and user information (name, profile, following/followers, tweets), with a maximum rate limit of 350 requests per hour. The third one, the Streaming API, is the most suited for data mining or analytic research allowing to retrieve a 1% filter of all tweets that users are actually carrying out, eventually using some filtering fields such as keywords, tags, users and geographic bounding box. Such rate limit can be raised asking Twitter for a gardenhose access, in order to receive a steady stream of tweets, very roughly, 10% of all public statuses. Note that these proportions are subject to unannounced adjustment as traffic volume varies.

Unlike Twitter, the Foursquare API allows to get all friends of an individual but does not consent, for reasons of privacy, to stalk a specified user. The only way to collect information about users activity is to select a set of venues in one or more specific regions, and download all the activities (check-ins) performed in those locations, with a rate limit of 5000 requests per hour. Both Twitter and Foursquare have severe lim-

itations to data retrieval, enabling to gather only a very partial subset of users activities.

Among many others geosocial network, Flickr poses the fewest limitations. In such on-line photo management system, practically all the valuable metadata such as tags and geolocation can be accessed by API programs. Anyway, some experiments carried out by the authors using the same query in different moments lead to retrieve slightly different results, leaving some uncertainty on the soundness of results. In some applications data can contain row geographic information such as coordinate (latitude and longitude) of the message generated by the mobile device, but also derived geographic information such as: coordinates bounding box, the place type, the place name or the street name. This information is produced by the social network application using the coordinated passed by the mobile device. The information that can be produced and retrieved changes depending on the system, the device, the privacy settings and so on.

1.4.4 Geographic uncertainty of CGI

Geosocial data can have several sources of uncertainty. In the following we describe some of them, we thought are the most important. Statistical analysis and future system developing have to consider this geographic uncertainty.

Uncertainty about precision. In this category we join issues related to data generation. The first one is information granularity: each point of the trajectory can have a different scale, sometimes coordinates of a specific place, sometimes a bounding box area. The second one is related to devices. A second case is that the precision can be modified by the social network system: in some applications (such as Foursquare, for example) the GPS data is used to infer higher level information, such as the address of a place and the passed coordinated are hidden. A third case is the location system used by the device. For example, in Figure 1.1 is possible to see a point located in Portoferraio, Elba island, on the left bottom corner of the trajectory. The user never went to the island. She set off the GPS for energy saving and the data was retrieved using the GSM antenna. Her mobile phone was attached to the Portoferraio antenna but she was physically on the cost, some kilometers north. There is no way to extract this information from the generated data.

Uncertainty about credibility. In some cases we can witness the presence of “spammer” users, which bombard the system of tweets and randomly

change the GPS coordinates relative to their geographic position to cheat the anti-spam system. In other cases people can voluntarily publish different location for fun or privacy reasons.

Uncertainty due to privacy settings. At the social network level, the geographic information can be filtered and modified for privacy issue in less detailed geographic level, despite the device transmits precise coordinates.

Uncertainty due to multiple data: tweet example. "Two very large forest fires in the mountains behind Funchal clouds of smoke covered the sun turn sunlight deep yellow ash coming down"[d]. The first piece of information is contained in the text in the toponym Funchal, more precisely the mountains behind the city of Funchal. The second information refers to a forest fire. Suppose the Tweet message itself has some coordinates originating from the source GPS device: this represents the third information. We can suppose that the user was sending the message safely far away from the forest fire, so it is possible to have several locations not spatially coincident (up to three in this example). Moreover, there is a certain level of uncertainty both in the definition of the mountain and in reverse geocoding of the toponym Funchal.

Uncertainty due to user and content location: photo example. Let us consider a person taking a picture with a camera or a smartphone with GPS integrated system. The person and the device coordinate overlap, while the subject of the photo coordinate has a distance from the camera. This distance could be considerable. Let us imagine that the content represented in the photo is the Mount Everest. The user with the camera is necessarily far from the mountain peak to include it into the photo. As shown in Figure 1.4 on the right side, the mountain and the camera could be consistently far one from the other.

These last two examples illustrate that there is a discrepancy between the location of the content (the registered device location at the time the message is sent or the photo is taken) and the geographic content contained in the message itself. The location of the device is not necessarily equal to the location of the reported content: they can overlap or be far away as in the examples. This inconsistency is not of technological nature, but will always include semantic aspects.

[d] Twitter posted by user Kevin bulmer on Fri Aug 13 2010 h20:21.

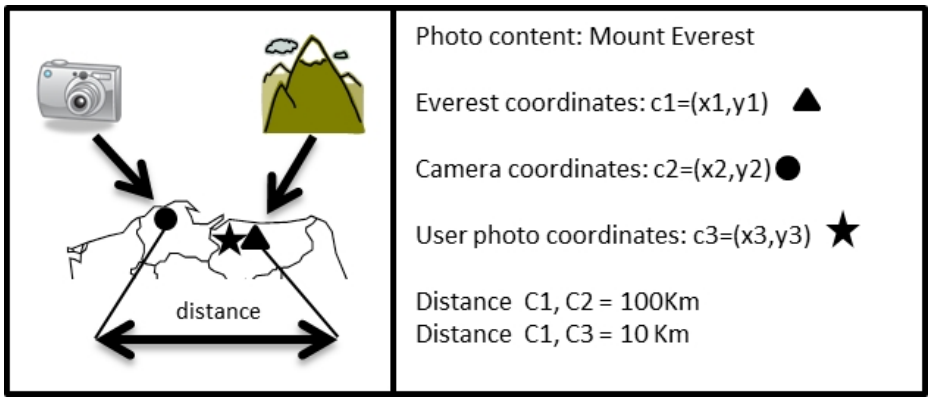

Figure 1.4 Comparison of device precision and semantic precision.

1.5 Open issues

Finally, we want to raise new key questions which we leave open for future research, that we believe will constitute interesting problems for the communities of Computer Scientists, Sociologists, Physicists, and Economists, for the years to come. Huge amounts of socially generated media resources on the Internet are a result of experience sharing by web communities. This fast growing media collection records our culture, society and environment, and provides opportunities to mine semantic and social knowledge of this world. Moreover, recent popularity of location-based social services, such as the Foursquare, Gowalla, Hot-Potato, etc., has generated huge amount of detailed location and event tags. It covers not only popular landmarks, but also obscure places, thus providing broad and wide coverage of locations in unprecedented scales. This large amount of information, often unstructured, opens a first research issue in the field of real time analysis of data flow. The research broadly covers several aspects such as: very large data repository in non-standard data structures; extracting semantic aggregation of tags; detecting places and event from unstructured text; finding automatic ways to link data from different sources; and many other examples. These topics are strictly linked to the developing of the so called Semantic Web and big companies as Google and Yahoo! are constantly researching and developing in these fields. Investigation on place and event semantics of geo-referenced tags, in addition to the representativeness, is a prerequisite to use geosocial data. A place tag is defined as a one that exhibits significant spatial

patterns, while an event tag refers to one that exhibits significant temporal patterns. Both definitions are vague and subject to some geographic region. For example, carnival may not be able to indicate any event, but will be very specific if only carnivals in New York City are considered. Analysing the spatial and temporal distribution of tags and identifying the distributions of events and places with relative geographic scales can be useful to many applications, such as image search, collection browsing, tag visualization and, of course, mobility analysis. Another open issue is the multilingualism of (geo-referenced) web media. Geo-Referenced media is, in facts, multilingual in nature. However, most systems take English as the sole processing language. This effectively excludes the media resources in other languages. The consequence is that the knowledge and patterns mined from geo-referenced media are biased towards English speaking countries and regions, while people are more comfortable to use their local language (also dialects and slang) to express with friends, especially in colloquial sentences such as the ones used in chats, SMS, status update or in stressful demanding situations like disasters or danger. The geographic locations of photos on the Internet have opened up a new host of research and application possibilities. As described in the photo example in Section 4, a spatial gap can exist between the GPS camera position and the position of the subject in the photo. Knowing the geographic orientation of photos, i.e., in which direction the cameras are pointing, will be useful to fill the gap. Though most cameras are not equipped with sensors to measure the orientation and inclination of the device, smart photos, with the iPhone and HTC Magic as prime examples, have started to embrace digital compass technologies. In addition to hardware sensors, software solutions to estimate photo orientation also exist, for example estimating the relative translation and orientation between photos, by leveraging the visual redundancy among photos. Till now, geographic orientation of photos are rarely available. Nevertheless, with the development of compass-equipped cameras and smartphones, such kind of metadata is expected to emerge in the near future. With the availability of photo orientation metadata, many compelling applications can be accomplished. For example, with the photo alignment information, visual summarization and browsing of photo collections can be adaptive to the user direction and perspective on the map. Moreover, 3D reconstruction of geolocation can be much more efficient.

1.6 Conclusion

We have discussed about mobility and geosocial networks, a very promising field of research nowadays, in which wide and multidisciplinary studies have been conducted in the last few years. We have seen how the interested in such topics is widely motivated by the interesting relationships that may reside between the social and mobility behaviour of humans: people move, they move with friends or relatives, they share experiences, they propagate information about new places to friends, and so on. Moreover, in the last years, it is clear how this process, supported by the large amount of online (geo)social networks and services, is extensively conducted on-line, in near real-time, with a clear social and participative trend. This kind of interactions and behaviours clearly produce massive amount of data about human actions, related to both social and mobility aspects, and opens the way for many interesting research challenges. However, despite the large interest and the large amount of data produced, we have seen how there is a clear disproportion between the results obtained so far, and the vast quantity and diversity of issues that are still open. We believe that the issues and peculiarities related to the data (availability, privacy, granularity, and so on) and the rapid explosion of the availability of new services and trends, are two clear reasons why it is still hard for the research in this direction to take off and to produce large and strong analytical results. The preliminary work conducted by many researcher so far is however very promising, and it seems clear that we are facing the start of a new era in the research on society and individual human behaviours.

1.7 Annotated bibliography

In order to complete this chapter, we present some works. We suggest to read them to deeper understand some ongoing research in the field of geosocial networks. Only the last three are related to trajectory, while the others deal with the geographic aspect of geosocial data. Warf and Sui (2010) work, in between geographic science and philosophy, mainly discusses how in practice neogeographers use geospatial technologies in multiple ways as opposite to conventional GIS. Spinsanti and Ostermann (2011) contribution describes the potential of CGI to dramatically change the traditional top-down, uni-directional communication pathway from official to public, via broadcasting media, in crisis man-

agement phases. Chorley at al. (2011), analysing a dataset composed by check-in data from Foursquare, reveal some individual characteristic of the cities. Cho et al. (2011) investigated the interaction of the persons social network structure and their mobility using datasets that capture human movements from Gowalla, Brightkite and phone location trace data. They tried to understand if friendships influence where people travel, or if it is more travelling that influences and creates social networks. The major contributions of Gaito et al. are the definition of the so-called geocommunity and the creation of a complex network-based methodology to extract geocommunities from GPS data applying clustering algorithm. Kisilevich et al. propose an approach for analysing trajectories of people, using geo-tagged photos collected from the photo-sharing site Flickr and a Wikipedia database of Point Of Interest (POI). In his article Purvess discuss the utilization of user generated content (UGC) as a data source for studying geographic questions, and propose two examples: the derivation of vernacular regions and trajectory analysis. Jankoski et al., in order to discover itineraries and preferences of landmarks in an urban context, aggregated geo-tagged photos downloaded from Flickr. They were able to find precise events that attracted attention of photographers. A spatial analysis of movement trajectories, led to interesting findings related to photographers itineraries. Lucchese et al. where able to extract, from photo published on Flickr, touristic point of interest in a city and provide automatically generated personalized recommendations.

References

Warf B. and Sui D., 2010. From GIS to neogeography: ontological implications and theories of truth. *Annals of GIS*, **16, 4**, 197–209.

Spinsanti L., Ostermann F., 2011. Retrieve volunteered geographic information for forest fire. *Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop, Milan, Italy, January 27-28*, **704**.

Chorley M. J., Colombo G. B., Williams M. J., Allen S. M., Whitaker R. M., 2011. Checking out checking in: observation on Foursquare usage patterns. *Finding Patterns of Human Behaviours in Network and Mobility Data NEMO, 2011.*

Cho E., Myers S. A., Leskovec J., 2011. Friendship and mobility: user movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011*, 1082–1090.

Gaito S., Rossi G. P., Zignani M., 2011. From mobility data to social attitudes: a complex network approach. *Finding Patterns of Human Behaviours in Network and Mobility Data NEMO, 2011.*

Kisilevich S., Keim D. A., Rokach L., 2010. A novel approach to mining travel sequences using collections of geotagged photos. *Proceedings of the 13th AGILE International Conference on Geographic Information Science, 2010.*

Purves R. S., 2011. Answering geographic questions with user generated content: experiences from the coal face. *Proceedings of the 27th annual ACM symposium on Computational geometry, New York, NY, USA, 2011*, 297–299.

Jankowski P., Andrienko N., Andrienko G., Kisilevich S., 2010. Discovering landmark preferences and movement patterns from photo postings *Transactions in GIS*, **14**, 1467-9671.